

Language Research Digital Repository

Jozef Mišutka



Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic

- Predstavenie projektov
 - LINDAT Clarin, META-NET META-SHARE
- Požiadavky technologického centra
- Technické riešenie, výhody a nevýhody
- Aktuálny stav

- Centre for Language Research Infrastructure
- “ ... umožnit národně distribuované, ale technologicky jednotné poskytování jazykových dat a souvisejících technologií všem zájemcům ... ”
- “V oblasti distribuce dat je cílem poskytovat službu repozitáře pro úschovu, licencování a poskytování dat v rámci celoevropské sítě ...”

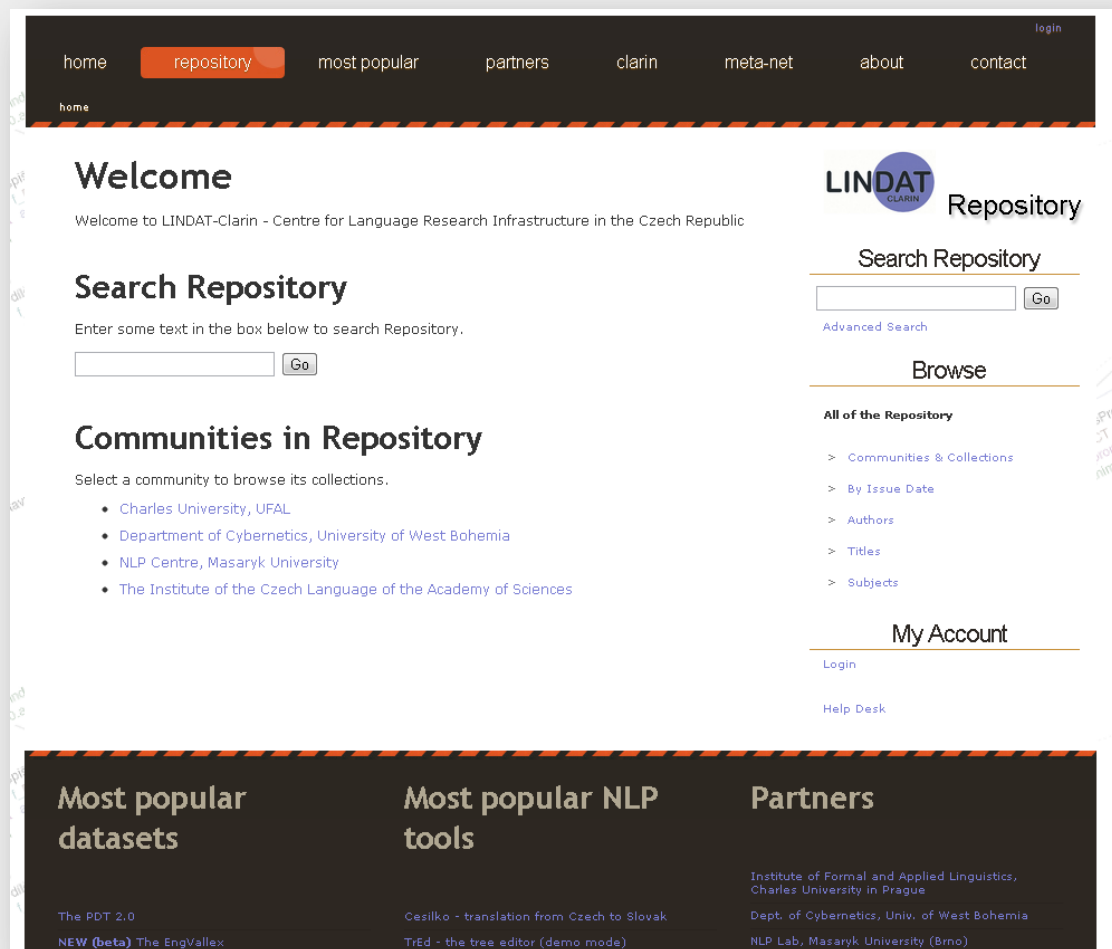
- Multilingual Europe Technology Alliance
- Vlastná implementácia DL
- “Repositories represent an evolution of the digital libraries paradigm towards open access, advanced search capabilities and large-scale distributed architectures. META-SHARE aims at providing such an open, distributed, secure, and interoperable infrastructure for the Language Technology domain.”

- Prístup a archivácia
- Pokročilé vyhľadávanie
- Metadáta, štrukturované (CMDI)
- Autentifikácia a autorizácia
- Licencie, IPR
- Export, import
- Persistentné identifikátory
- Štatistiky
- Data Seal of Approval

- Prístup a archivácia
- Pokročilé vyhľadávanie
- Metadáta, štrukturované (CMDI)
- Autentifikácia a autorizácia
- Licencie, IPR
- Export, import
- Persistentné identifikátory
- Štatistiky
- Data Seal of Approval

Reklama na DSPACE? Áno, ale ...

- DSpace 1.6.2 - ostrá prevádzka
 - prechod na 1.8.2
- 24/7 - HW+SW vďaka virtualizácii a redundancii
 - v extrémnom prípade môže tiež zlyhať
- Technologický štandard (viac-menej :)
 - apache, tomcat, git, redmine, postgresql, makefiles, mutli instance deploy, simple qa, vm+data bckp/restore, monitoring, EduID/czTestFed, google analytics, webmaster



The screenshot shows the LINDAT-Clarín Repository website. The top navigation bar includes links for home, repository (highlighted), most popular, partners, clarin, meta-net, about, and contact. A login link is also present in the top right corner.

The main content area is divided into several sections:

- Welcome:** A message welcoming users to the LINDAT-Clarín - Centre for Language Research Infrastructure in the Czech Republic.
- Search Repository:** A section with a search input field and a 'Go' button. Below the input field is a link for 'Advanced Search'.
- Browse:** A section with a list of categories to browse: 'All of the Repository', 'Communities & Collections', 'By Issue Date', 'Authors', 'Titles', and 'Subjects'.
- My Account:** A section with links for 'Login' and 'Help Desk'.
- Most popular datasets:** A section listing 'The PDT 2.0' and 'NEW (beta) The EngVallex'.
- Most popular NLP tools:** A section listing 'Cesilko - translation from Czech to Slovak' and 'TrEd - the tree editor (demo mode)'.
- Partners:** A section listing 'Institute of Formal and Applied Linguistics, Charles University in Prague', 'Dept. of Cybernetics, Univ. of West Bohemia', and 'NLP Lab, Masaryk University (Brno)'.

- Hlavná stránka projektu

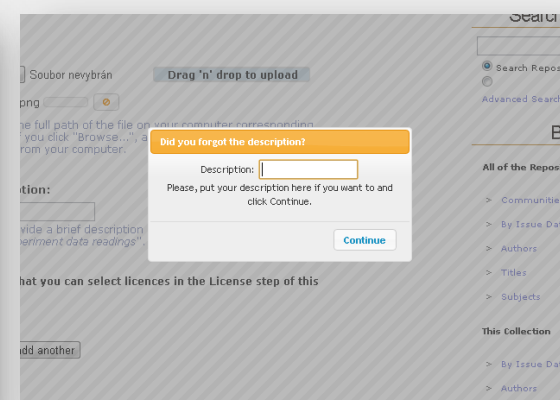
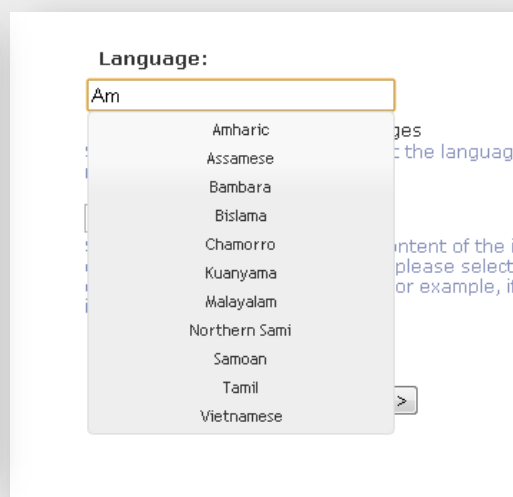
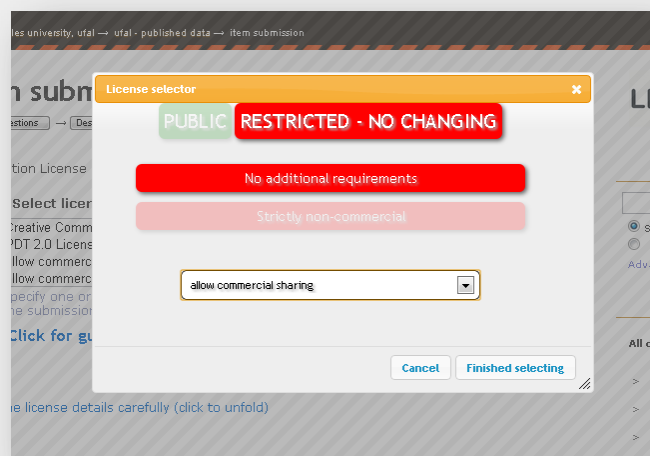
<https://ufal-point.mff.cuni.cz/xmlui/>

- Použiteľné out of the box (pri minor verzii 2+)
- Veľká komunita
- Dspace z veľkej časti plní požiadavky na TC
- Zlepšujúca sa dokumentácia
- Zlepšujúca sa podpora administrátorského rozhrania
- Základné veci dobre konfigurovateľné
- Dodržiavanie štandardov
- Zlepšujúca sa podpora real world nasadenia

- Upgrade proces
- Pomalá “krivka učenia”
- Niektoré základné a väčšina netriviálnych zmien je časovo aj technicky náročná
 - výhodné robiť veci mimo DSpace
- Niektoré procesy sú pevné
 - DSpace si myslí, že vie najlepšie čo chceme
- Neprehľadný databázový model
 - Stále nedostatočné administrátorské rozhranie
- Ladenie, kompilácia a redeploy
- Iba triviálna podpora licencií
- Užívateľská a programátorská prívetivosť GUI

- Vyhľadávanie
 - základná konfigurácia
- Metadáta
 - hierarchické metadát
- Shibboleth
 - problémy
- Licencie
- Export/import a PIDs
- Nutný “programátor”

- Uživatel'sky “prívětivější”



- Uživatel'sky "prívětivější"

File:

Soubor nevybrán

Drag 'n' drop to upload



Picture3.png

Please enter the full path of the file on your computer corresponding to your item. If you click "Browse...", a new window will allow you to select the file from your computer.

sentence information structure and coreference relations are annotated at the semantic level. PDT 2.0 is based on the long-standing Praguian linguistic tradition, adapted for the current Computational Linguistics research needs. The corpus itself uses the latest annotation technology. Software tools for corpus search, annotation and language analysis are included. Extensive documentation (in English) is provided as well.

Handle: <http://hdl.handle.net/11858/00-097C-0000-0001-8098-5>

Date: 2006-07-21

Files in this item

Files	Size	Format	View	Description	License
a-man-cz.pdf	4.238Mb	PDF	View/ Open	Manual for Analytical Annotation (Czech)	ACA
a-man-en.pdf	1.639Mb	PDF	View/ Open	Manual for Analytical Annotation (English)	ACA
m-man-en.pdf	351.4Kb	PDF	View/ Open	Manual for Morphological Annotation (English)	ACA
pdt2.zip	229.1Mb	Unknown	View/	Prague Dependency Treebank 2.0 data	ACA

Files in this item

Files	Size	Format	View	Description	License
pdt2sample.zip	549.2Kb	Unknown	View/ Open	A sample of the full PDT 2.0 data	CC

This item appears in the following Collection(s)

> [Authors](#)

> [Titles](#)

> [Subjects](#)

Distributed under Creative Commons (allow commercial sharing and changing)

- Uživatel'sky "prívětivější"

UFAL Licence agreement page!

If you want to see the content you have to agree the licence below

- [PDT 2.0 License](#)

SIGNER: Jozef Mišutka

ID of signer: jmisutka@gmail.com

ITEM handle: 11858/00-097C-0000-0001-CC75-3

BITSTREAM: [173](#)

LICENSE:

By following this link I, SIGNER with ID, agree with the LICENSE restricting the usage of the ITEM's BITSTREAM.

This bitstream is restricted

You do not have the credentials to access the restricted bitstream **stopwords_en.txt**.

Control Panel

[Java Information](#) [Configuration](#) [System-wide Alerts](#) [Harvesting](#) [Current Activity](#) [User Logins](#)

jmisutka@gmail.com @ 2012-05-08 11:51:41,68	
host	ufal-point.mff.cuni.cz:8080
cache-control	max-age
user-agent	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/536.9 (KHTML, like Gecko) Chrome/20.0.1120.0 Safari/536.9
accept	text/html,application/xhtml+xml,application/xml;q
referer	https://cas.cuni.cz/idp/Authn/RemoteUser?ticket
accept-encoding	gzip,deflate,sdch
accept-language	cs,en-GB;q
accept-charset	windows-1250,utf-8;q
cookie	JSESSIONID
shib-session-id	21a847fb53cdadbf32a9d25593cd89bf

- Uživatel'sky "prívětivější"

Edit Item

[Item Status](#)
[Item Bitstreams](#)
[Item Metadata](#)
[View Item](#)

Add new metadata

Name:

Value:

PLEASE enter data in a way. You are responsible for entering the data in a way. If you are not sure what the format is, please do NOT make changes.

[Add new metadata](#)

Metadata
[Update](#) [Return](#)

Remove	Name	Value	Language
<input type="checkbox"/>	dc.contributor.author	Jozef, Misutka	<input type="text"/>

Distribution medium:

Specifies the format used for the delivery of the resource.

ResourceInfo ValidationInfo

Validation status:

The validation status of the resource; please, use "yes" even for partially validated resources. Has the resource been validated in some form (content and/or format validation)?

ResourceInfo ResourceCreationInfo

Project title:

The full title of the project that led to the creation of the resource or tool/service.

☐ #1-1
☐ #2-2

Funding type:

The source of the funding of the data creation - possible values are EU, Own, National, 3rd party.

☐ #1-1
☐ #2-2

[Remove selected component values](#)
[Add component](#)

- DSpace 1.8
- EU sign-on
- Lepšia administrácia
- CMDI
- Profesionálne vyhľadávanie
- Import / export

- Problémový bottom-up model
- Licencie
- Otvorené metadáta
- Výhody Dspace



Shibboleth®

 **META**



ÚFAL

